Big Data et documentation

Jean-Michel Salaün
ENS de Lyon

Journée d'étude "Mieux accompagner la recherche : réalités d'aujourd'hui et perspectives pour les fonctions d'IST"
30 septembre 2015

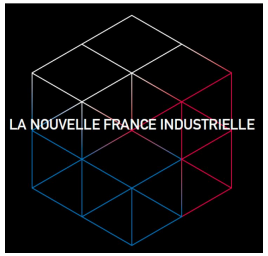
Chiffre d'affaires

« Les activités liées au Big Data ont représenté en France environ 1,5 milliards d'euros cette année, et atteindront près de 9 milliards d'ici 2020 (recouvrant les dépenses en logiciels, services, et dépenses internes des entreprises). »

La nouvelle France industrielle, Big Data – Feuille de route, Paul HERMELIN (PDG de CapGemini) et François BOURDONCLE (Président de FB&Cie). 2 juillet 2014

Emplois

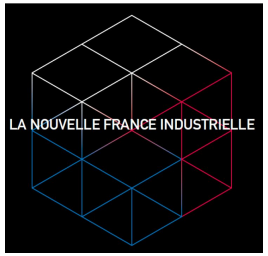
« En termes d'emplois, l'enjeu du Big Data est de créer ou consolider de l'ordre de 137 000 emplois, soit directement dans l'industrie informatique, soit dans les fonctions technologiques au sein des entreprises, soit dans les fonctions métiers (par exemple les fonctions de vente) des entreprises. »



La nouvelle France industrielle, Big Data – Feuille de route, Paul HERMELIN (PDG de CapGemini) et François BOURDONCLE (Président de FB&Cie). 2 juillet 2014

Activités

« A l'instar de ce qui s'est passé pour la presse, le commerce de proximité, le tourisme ou encore l'industrie musicale, ces sociétés [du numérique] tentent actuellement d'utiliser [les données sur leurs clients] pour modifier profondément le paysage concurrentiel, et en particulier la relation client et la chaîne de valeur.
C'est ainsi que des secteurs comme l'assurance, la banque, la grande distribution, le crédit à la consommation, l'industrie automobile, les «utilities », l'énergie, et bien d'autres encore risquent de voir leur positionnement dans la chaîne de valeur réduite progressivement au rôle de soustraitant « technique », avec une très forte pression sur les marges et un risque potentiellement létal pour les plus faibles. »



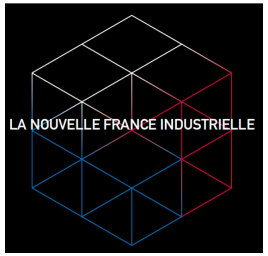
La nouvelle France industrielle, Big Data – Feuille de route, Paul HERMELIN (PDG de CapGemini) et François BOURDONCLE (Président de FB&Cie). 2 juillet 2014

Des « data-scientists »

1. Actions de nature à développer l'écosystème Big Data en France

Action 1 Formation de « data scientists »
... p.6

Le consensus aujourd'hui est de définir le data scientist à l'intersection de trois domaines d'expertise : (i) l'informatique, (ii) les statistiques et les mathématiques, et (iii) les connaissances métier. p.7



La nouvelle France industrielle, Big Data – Feuille de route, Paul HERMELIN (PDG de CapGemini) et François BOURDONCLE (Président de FB&Cie). 2 juillet 2014

Le « big data »

Ce sont les petites miettes de données que vous laissez derrière vous quand vous vous déplacez sur terre.
Ce que ces miettes racontent, c'est l'histoire de votre vie. Elles disent ce que vous avez choisi de faire. C'est très différent de ce que vous mettez sur Facebook. Ce que vous mettez sur Facebook, c'est ce que vous voudriez dire aux gens, rédigé selon les normes d'aujourd'hui. (...)
Si je peux connaitre certains de vos comportements, je peux inférer le reste juste en vous comparant avec la foule de ceux qui vous ressemblent. (…)

Alex Pentland in Reinventing Society In The Wake Of Big Data. Edge, août 30, 2012.

Deux oublis

- La compétence d'organisation de l'information dans les compétences-métiers

- Les données ne sont pas que des traces

Quelles connaissances métiers ?

- Deux « connaissances métiers » confondues :
  - Les domaines d'application : banques, voyage, tourisme, ecommerce, publicité, transport, édition, automobile, etc.
  - L'architecture de l'information : conception de systèmes d'information pour les utilisateurs.

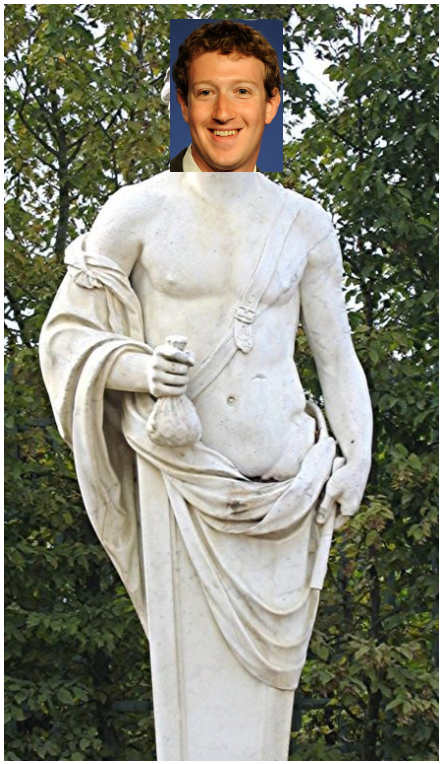
Big, big, big...

« La conception et la mise en œuvre d'une infrastructure d'intermédiation se fait selon l'organisation de trois types de données différentes qui suivent des logiques particulières et se superposent : les données brutes ou primaires (*big data*), les métadonnées ou données d'index (*big index*) et les données de profilage des utilisateurs (*big user*). » S. Frénot

Big data : Wikipédia, bibliothèques, BDD publiques, open data...
Big index : Google, Booking.com...
Big users : Facebook, Linkedin, Uber...

Les trois « big » nécessitent des infrastructures et compétences informatiques différentes.

Prométhée et Hermès

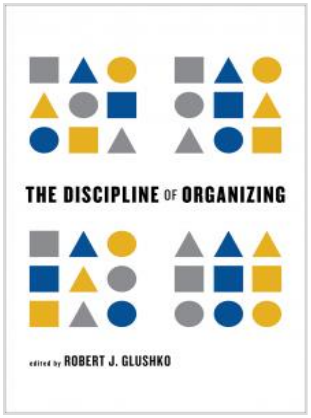


Prométhée enchaîné

« Je suis prêt à sacrifier tout cela parce que je ne peux, en mon âme et conscience, laisser le gouvernement américain détruire la vie privée, la liberté d'Internet et les libertés essentielles des gens du monde entier avec ce système énorme de surveillance qu'il est en train de bâtir secrètement » (2013)

Revisiter les sciences de l'information

« Organiser, c'est ouvrir des possibilités
en imposant volontairement
un **ordre** et une **structure**. »
Robert J. Glushko
Prix ASIS&T du livre de l'année 2014
Nouvelle édition augmentée en 2015



Les trois moments de l'organisation

1. Le stockage
2. La logique de l'organisation
3. La présentation

Dans le monde de l'information

Moments	En bibliothèque	Avec un moteur de recherche
Moment 1 : Stockage	Acquisition et	Crawling
Moment 2 : Logique d'organisation	Catalogage et description	Indexation
Moment 3 : Présentation	catalogage et résultats	s

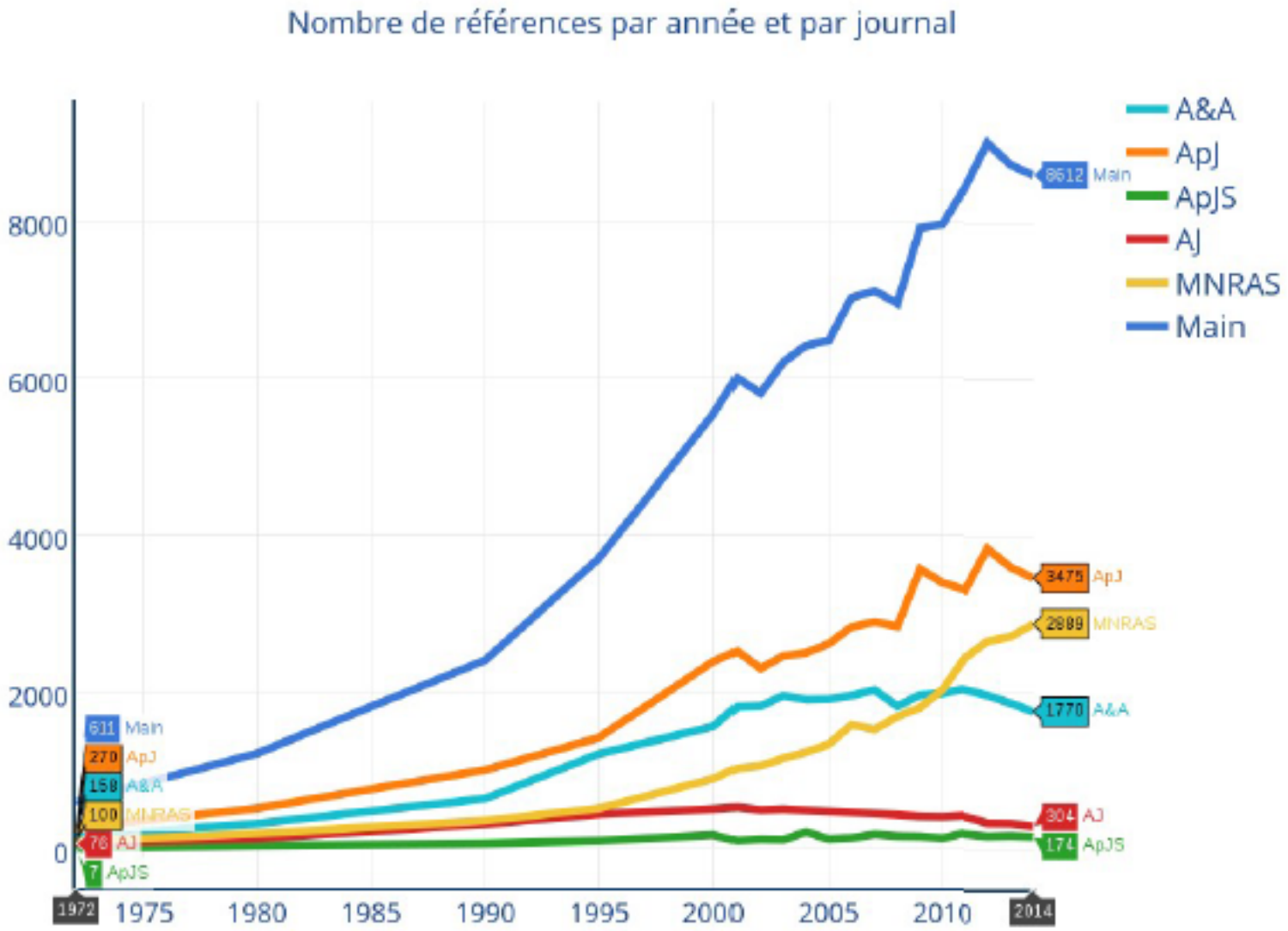
Salle des catalogues bibliothèque du Congrès, début du 20e siècle (big index)

Centre de données de Google, début 21e siècle (big index)

Les cathédrales du savoir

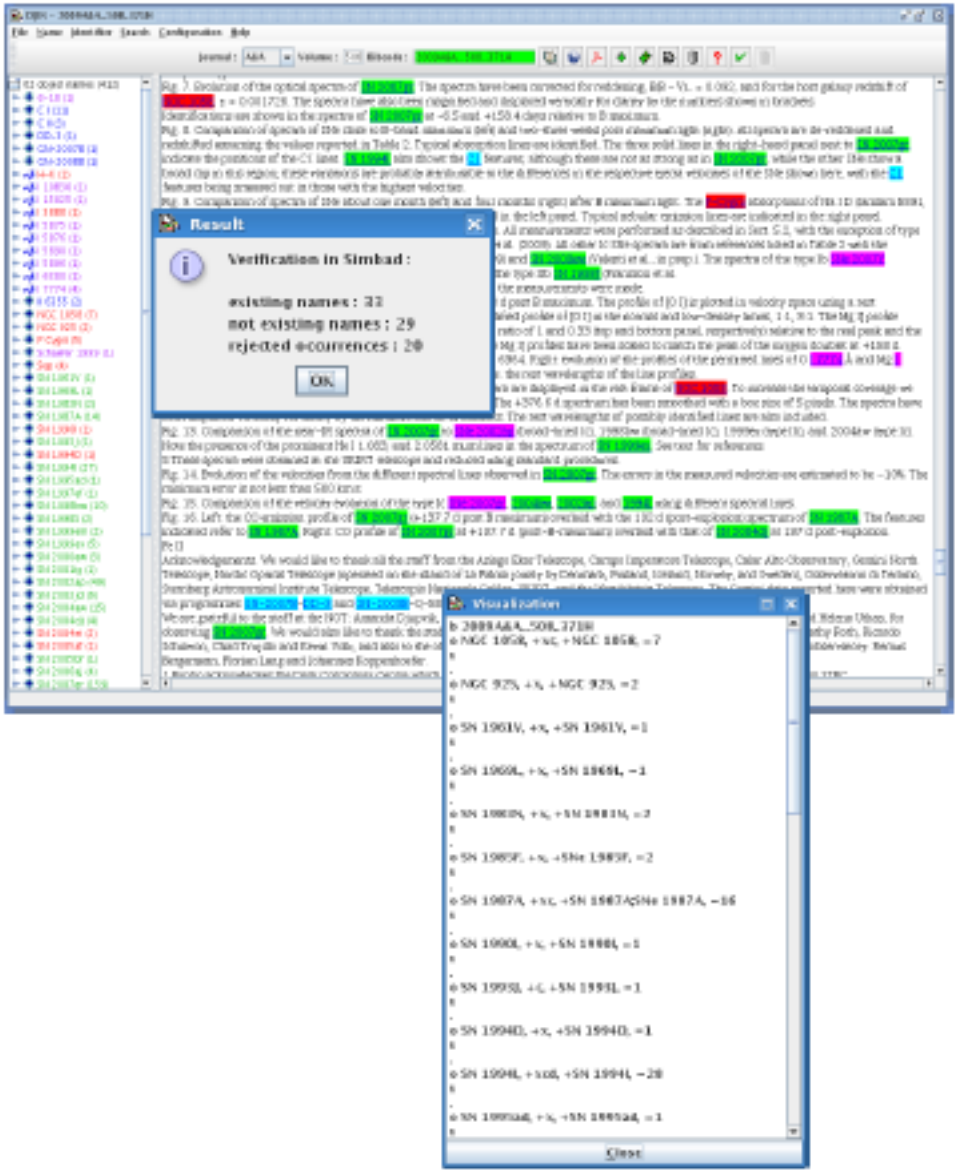
Boston Public Library
Photo Père Ubu

Les usines de lecture


Nombre de références par année et par journal
A&A
ApJ
ApJS
AJ
MNRAS
Main
8000
6000
4000
2000
0
8612 Main
3475 ApJ
2889 MNRAS
1770 A&A
304 AJ
174 ApJS
611 Main
270 ApJ
150 A&A
100 MNRAS
76 AJ
7 ApJS
1972
1975
1980
1985
1990
1995
2000
2005
2010
2014

La construction du contenu

DJIN (Detection in Journals of Identifiers and Names)



Ce programme developpé au CDS et opérationnel depuis 2008 permet de reconnaître semi-automatique des noms d'objets astronomiques à partir du texte complet accessible en ligne.
Ce logiciel est basé sur le dictionnaire de nomenclature des objets célestes.

Une validation par des documentalistes est nécessaire pour :

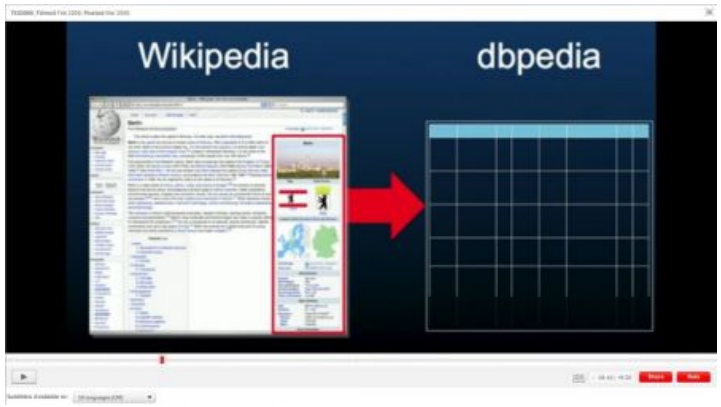
- Valider ou non ces objets
- Rechercher les objets non détectés automatiquement
- Rechercher toute information complémentaire attachée à ces objets

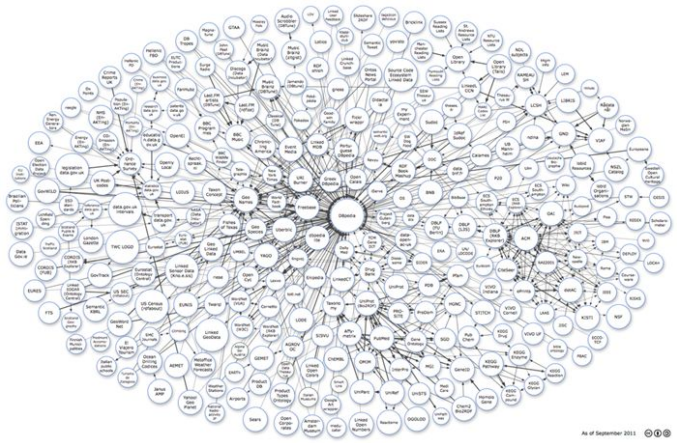
Échanges avec un astronome référent en cas d'objets complexes.

Valeurs ajoutées : occurence, positions

Faciliter le travail des documentalistes et mettre en avant leur expertise scientifique.

L'exploitation du web de données





Big data → **Big indexes**

L'exploitation des traces

ISI Web of KNOWLEDGE
Transforming Research

7 millions d'utilisateurs

ScienceDirect

academia.edu

HAL

Big data → **Big users**

Du document à la donnée
Transformation de notre relation au savoir

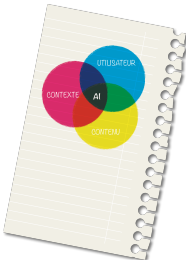
- 19e-20e : Document scientifique = article de revue et livre

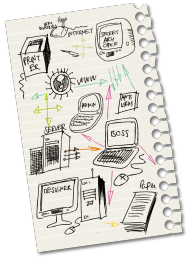
» Transmettre et prouver

- 21e : ressources scientifiques
  - Documents numériques (hyper-) (Web 1)
    - ArXive, ScienceDirect…
    - Gallica, Europeana, Google-Book…
  - Néodocuments (Web 2 - traces)
    - Blogues
    - Wikipédia
    - Réseaux sociaux
    - Sites dynamiques, etc.
  - Données (Web 3 – open data)

» Partager et coconstruire ?

Vers des architectes de l'information





L'architecte de l'information a six compétences principales